Multivariate Behavioral Research, 34 (2), 181-197

Structural Equation Modeling with Small Samples: Test Statistics

Peter M. Bentler
University of California, Los Angeles

Ke-Hai Yuan
University of North Texas

Structural equation modeling is a well-known technique for studying relationships among multivariate data. In practice, high dimensional nonnormal data with small to medium sample sizes are very common, and large sample theory, on which almost all modeling statistics are based, cannot be invoked for model evaluation with test statistics. The most natural method for nonnormal data, the asymptotically distribution free procedure, is not defined when the sample size is less than the number of nonduplicated elements in the sample covariance. Since normal theory maximum likelihood estimation remains defined for intermediate to small sample size, it may be invoked but with the probable consequence of distorted performance in model evaluation. This article studies the small sample behavior of several test statistics that are based on maximum likelihood estimator, but are designed to perform better with nonnormal data. We aim to identify statistics that work reasonably well for a range of small sample sizes and distribution conditions. Monte Carlo results indicate that Yuan and Bentler's recently proposed *F*-statistic performs satisfactorily.

Introduction

Structural equation modeling, especially its special case of covariance structure analysis, has been used extensively in the psychological, social, and behavioral sciences. Although there are many aspects to modeling, such as parameter estimation, model testing, and evaluating the size and significance of specific parameters, typically model evaluation by a goodness of fit test statistic represents the most critical step in modeling. After all, there is not much point to worrying about specific parameters in the context of a model that is not consistent with the data. This article addresses the problem of model evaluation when sample size is small and data may be nonnormal. See for example, Browne and Arminger (1995) for a review of the statistical theory, Bentler and Dudgeon (1996) for a discussion of consequences of violation of assumptions, and Austin and Calderón (1996) for a guide to the literature.

Since Jöreskog's (1969) emphasis, the most widely utilized test statistic in this field is the classical likelihood ratio statistic based on normal theory

This research was supported in part by the National Institute on Drug Abuse, grants DA01070 and DA00017.

maximum likelihood (ML) estimation. An advantage of ML is that it can be applied even when sample size is quite small, perhaps only slightly larger than the number of variables in the analysis, but an important disadvantage is that it can yield quite distorted conclusions about model adequacy under violation of the requisite assumption of multivariate normality. In fact, psychological data often are nonnormal. For example, Micceri (1989, p. 156) reported that "An investigation of the distributional characteristics of 440 large-sample achievement and psychometric measures found all to be significantly nonnormal at the alpha .01 significance level. Several classes of contamination were found...the underlying tenets of normality-assuming statistics appear fallacious for these commonly used types of data." This means that an alternative method that does not invoke the normality assumption would be ideal. The most elegant such method is the asymptotically distribution free (ADF) method and its associated test statistic (Browne, 1984). Unfortunately, this classic method needs medium to large sample sizes to get stable estimators, and unreasonably large sample sizes to make the ADF test statistic behave as a nominal chi-square variate. See Hu, Bentler, and Kano (1992), Muthén and Kaplan (1992), and Curran, West, and Finch (1996) for details. Although Yuan and Bentler (1997a) developed a finite sample correction to the ADF statistic that permits ADF testing in intermediate sample sizes, it too is limited by a fundamental lower-bound sample size required by this test. When sample size is less than the number of nonduplicated elements in the sample covariance matrix, both Browne's ADF procedure and the Yuan-Bentler modification cannot be performed at all since a critical weight matrix that must be inverted necessarily is singular. This means that an appropriate procedure for sample sizes smaller than this critical value will require the use of other methods.

The potentially most promising procedure for this situation is a relatively unknown residual-based test procedure developed by Browne (1984). It can be applied to any consistent estimators including the ML estimator, even when data are not normally distributed. Under its theoretical conditions, it remains asymptotically chi-square distributed. However, Yuan and Bentler (1998) showed that the residual-based ADF statistic, like the classical ADF statistic, requires a very large sample size to give reliable inference. An alternative methodology is the rescaled statistic of Satorra and Bentler (SB, 1988, 1994), which multiplies the ML test statistic by a correction factor that depends on the data and the model. Although this method has been shown to work very well in practice (see e.g., Hu et al., 1992; Curran et al., 1996), an unsatisfactory aspect of the SB rescaled statistic is that its theoretical null distribution is generally unknown for a nonnormal data set. To obtain a wider variety of methods, Yuan and Bentler (1998) proposed several new statistics

with known asymptotic distributions. Empirical studies also have indicated that some of these new statistics also behave well for medium sized samples. However, the behavior of these new statistics has not been studied for small sample sizes. In fact, there is no literature addressing the behavior of these various test statistics in samples that are smaller than the number of nonduplicated elements of the covariance matrix, which is typical in many applications. Clearly this situation violates the assumption of asymptotic sample sizes that arises with all extant test statistics because of the nonlinearity of typical covariance structures.

When theoretical analysis does not yield a clear choice among methods, empirical Monte Carlo study is needed to describe the behavior of the method. Here we are interested in the small sample behavior of various test statistics under conditions of nonnormality, because such conditions are especially relevant to practical data sets. With the aim for finding reliable statistics for this situation, we will study the following statistics: The normal theory based likelihood ratio statistic T_{ML}, the Satorra-Bentler rescaled statistic T_{SB}, the Yuan and Bentler version of residual based ADF statistic T_{YB}, and an F-statistic derived from the residual-based ADF statistic. For comparison purposes, we also report the performance of the residual-based ADF statistic T_B. We will study these statistics for both normal and nonnormal data. These statistics and the designed conditions will be introduced in the following section. Results of our simulation study will be presented in the section following that. Conclusions and remarks will be given at the end of this article.

Statistics and Designed Conditions

Let $X_i = \left(x_{i1}, \ldots, x_{ip}\right)'$, $i = 1, \ldots, N = n+1$ be a sample from $X = \left(x_1, \ldots, x_p\right)'$, with sample mean $\overline{X}$ and sample covariance S. For a covariance structure $\Sigma = \Sigma(\theta_0)$, the estimate $\hat{\theta}$ of the unknown parameter θ_0 can be obtained by minimizing

$$F_{ML}(\theta) = tr\left[S\Sigma^{-1}(\theta)\right] - \log|S\Sigma^{-1}(\theta)| - p. \tag{1}$$

Under the assumption of multivariate normality and the null hypothesis, $T_{ML} = nF_{ML}\left(\hat{\theta}\right)$ is asymptotically distributed as $\chi^2_{p^*-q}$, where $p^* = p(p+1)/2$ and q is the number of unknown parameters in θ_0. When the normality assumption does not hold, we can still estimate the unknown parameter θ_0 by minimizing the function $F_{ML}(\theta)$, but T_{ML} will generally not

approach $\chi^2_{p^*-q}$ anymore. Realizing that real data sets in practice seldom follow normal distributions, Browne (1984) proposed a test statistic which does not require any specific distribution assumptions. Let vech(·) be an operator which transforms a symmetric matrix into a vector by stacking the nonduplicated elements of the matrix, s = vech(S), $\sigma(\theta)$ = vech[$\Sigma(\theta)$], and denote the $p^* \times q$ Jacobian matrix corresponding to $\sigma(\theta)$ as $\dot{\sigma}(\theta)$. Then there exists a full column rank $p^* \times (p^* - q)$ matrix $\dot{\sigma}_c(\theta)$ whose columns are orthogonal to those of $\dot{\sigma}(\theta)$. Let Y_i = vech[$(X_i - \bar{X})(X_i - \bar{X})'$], and S_Y be the corresponding sample covariance matrix of Y_i. Then S_Y is a consistent estimate of Γ = cov{vech[$(X - \mu)(X - \mu)'$]}, where $\mu = E(X)$. For the estimate $\hat{\theta}$, the residual-based test statistic given by Browne (1984) is

$$T_B(\hat{\vartheta}) = n\hat{e}'\dot{\sigma}_c(\hat{\vartheta})\left[\dot{\sigma}_c'(\hat{\vartheta})S_Y\dot{\sigma}_c(\hat{\vartheta})\right]^{-1}\dot{\sigma}_c'(\hat{\vartheta})\hat{e}, \tag{2}$$

where $\hat{e} = s - \sigma(\hat{\theta})$ is the discrepancy between the data and the model estimated by any consistent estimator. We use the ML estimator for T_B.

Let $W = 2^{-1} D_p'(\Sigma^{-1} \otimes \Sigma^{-1})D_p$, where D_p is the $p^2 \times p^*$ duplication matrix as defined in Magnus and Neudecker (1988, p. 49), $\dot{\sigma} = \dot{\sigma}(\theta_0)$, and

$$U = W - W\dot{\sigma}(\dot{\sigma}'W\dot{\sigma})^{-1}\dot{\sigma}'W. \tag{3}$$

Then

$$T_{ML} \xrightarrow{\mathcal{L}} \sum_{j=1}^{p^*-q} \tau_j \chi^2_{j1},$$

where τ_j are the nonzero eigenvalues of $U\Gamma$ and χ^2_{j1} are independent chi-square distributions with degree of freedom 1. So, for a general nonnormal distribution, even the mean of the asymptotic distribution of T_{ML} does not match that of the nominal $\chi^2_{p^*-q}$. Since

$$E\left(\sum_{j=1}^{p^*-q} \tau_j \chi^2_{j1}\right) = \sum_{j=1}^{p^*-q} \tau_j \tag{4}$$

and $tr(\hat{U}S_Y)$ is a consistent estimate of the right hand side of Equation 4, Satorra and Bentler (1988, 1994) proposed a statistic

$$T_{SB} = \frac{p^* - q}{tr(\hat{U}S_Y)} T_{ML}, \tag{5}$$

where $\hat{U}$ is a consistent estimate of U. Even though the asymptotic distribution of T_{SB} is still not $\chi^2_{p^*-q}$, it at least matches the first moment of the nominal chi-square distribution. Existing empirical studies support this statistic under variety of conditions.

Since the statistic T_B requires extremely large sample size to be reliable, and the asymptotic distribution of T_{SB} is generally unknown, Yuan and Bentler (1998) proposed several new statistics. In the regression literature, cross-products of model residuals are regularly used for estimating asymptotic covariances and standard errors. For a consistent $\hat{\theta}$, we can estimate Γ by

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{N}\left[Y_i - \sigma\left(\hat{\theta}\right)\right]\left[Y_i - \sigma\left(\hat{\theta}\right)\right]' = S_Y + \frac{N}{n}\left[\bar{Y} - \sigma\left(\hat{\theta}\right)\right]\left[\bar{Y} - \sigma\left(\hat{\theta}\right)\right]'.$$

Replacing S_Y in Equation 2 by this estimate, Yuan and Bentler obtained the following statistic

$$T_{YB}\left(\hat{\theta}\right) = T_B\left(\hat{\theta}\right) / \left[1 + NT_B\left(\hat{\theta}\right)/n^2\right]. \tag{6}$$

Since $T_{YB} < T_B$, the problem of over rejection with T_B can be possibly resolved by T_{YB}. Equation 6 also implies that T_B and T_{YB} are asymptotically equivalent, so both T_B and T_{YB} asymptotically follow $\chi^2_{p^*-q}$.

For testing the hypothesis $A\mu = b$, the well-known Hotelling's T^2 statistic is

$$T^2 = N\left(A\bar{X} - b\right)'(ASA')^{-1}\left(A\bar{X} - b\right). \tag{7}$$

Rewriting Equation 2 as

$$T_B\left(\hat{\theta}\right) = n\left[\dot{\sigma}_c'\left(\hat{\theta}\right)\hat{e}\right]'\left\{\dot{\sigma}_c'\left(\hat{\theta}\right)S_Y\dot{\sigma}_c\left(\hat{\theta}\right)\right\}^{-1}\left[\dot{\sigma}_c'\left(\hat{\theta}\right)\hat{e}\right], \tag{8}$$

and comparing Equation 8 with Equation 7, we will find that Equation 8 corresponds to testing the null hypothesis $\dot{\sigma}_c'(\theta_0)\left[\sigma(\theta_0) - \sigma\right] = 0$. Based on

such an observation, Yuan and Bentler (1998, in press) proposed to use a Hotelling's T^2 distribution to approximate that of T_B instead of a chi-square, leading to an F-statistic which is given by

$$T_F = \left[N - \left(p^* - q\right)\right] T_B \,/\, \left[(N-1)\left(p^* - q\right)\right], \tag{9}$$

and is referred to an F-distribution with degrees of freedom $p^* - q$ and $N - (p^* - q)$. Note that the statistics T_{YB} and T_F are asymptotically equivalent with T_B. However, as we shall see, it is quite likely that their performances will differ when applied to finite samples.

It is obvious that the $p^* - q$ square matrix $\left[\dot{\sigma}_c'\left(\hat{\theta}\right) S_Y \dot{\sigma}_c\left(\hat{\theta}\right)\right]$ has to be invertable in order for T_B, and consequently for T_{YB} and T_F, to be defined. Since the rank of S_Y is the minimum of p^* and $N - 1$, this means that the sample size has to be at least $N \geq p^* - q + 1$. This requirement can be better understood through Hotelling's T^2 distribution. When using the T^2 for testing a population mean with dimension $r = p^* - q$, the statistic $F = (N-r)T^2 \,/\, \left[r(N-1)\right]$ is compared to an F-distribution with degrees of freedom r and $N - r$. So $p^* - q + 1$ is the smallest possible sample size that yields a positive degree of freedom $N - r$. Also, it seems that the statistic T_{SB} can still be computed for any small sample size, even one smaller than $p^* - q + 1$. However, the matrix $\hat{U}S_Y$ needs to have a rank $p^* - q$ in order for the rescaling factor $tr\left(\hat{U}S_Y\right) / \left(p^* - q\right)$ to make sense; this is because the scaling factor represents the average nonzero eigenvalue of the given matrix product. Although it can be shown with Equation 3 that rank $\left(\hat{U}S_Y\right) = p^* - q$ is equivalent to the requirement that $\left[\dot{\sigma}_c'\left(\hat{\theta}\right) S_Y \dot{\sigma}_c\left(\hat{\theta}\right)\right]$ be full rank, it nonetheless is possible to use T_{SB} even when rank $\left(\hat{U}S_Y\right)$ is less than $p^* - q$. However, the effect of such a practice is not known. We will investigate this empirically.

A final consideration is whether any of the statistics we study can be expected to be robust to violation of its conditions. Even though the asymptotic distributions of T_{ML} and T_{SB} are generally unknown for nonnormal data, conditions exist for them to asymptotically follow $\chi^2_{p^*-q}$ (e.g., Amemiya & Anderson, 1990; Browne & Shapiro, 1988; Satorra & Bentler, 1990). These conditions involve independencies among latent generating variates as sample size increases arbitrarily, and, unfortunately, there is no effective way of verifying these conditions in practice. Further, since our interest is in the small sample behavior of the statistics, asymptotic robustness theory may not really be relevant here. Nonetheless, it is still of interest to know if there are any differences in performance of the statistics under a variety of conditions.

We will use the following confirmatory factor model

$$X = \Lambda f + e \quad \text{with } cov(X) = \Lambda\Phi\Lambda' + \Psi, \tag{10}$$

and

$$\Lambda = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1.0 & .30 & .40 \\ .30 & 1.0 & .50 \\ .40 & .50 & 1.0 \end{pmatrix}, \tag{11}$$

where $\lambda' = (.70, .70, .75, .80, .80)$ and 0 is a vector of five zeros. The Ψ is a diagonal matrix which makes Σ a correlation matrix. In order for this model

Table 1
Designed Conditions[a]

I	$X = \Lambda f + e$, $f \sim N(0,\Phi)$, $e \sim N(0,\Psi)$
II	$X = (\Lambda f + e)/r$, $f \sim N(0,\Phi)$, $e \sim N(0,\Psi)$, $r \sim \sqrt{\chi^2_5/3}$
III	$X = (\Lambda f + e)/r$, $f \sim N(0,\Phi)$, $e \sim$ Lognormal$(0,\Psi)$, $r \sim \sqrt{\chi^2_5/3}$
IV	$X = (\Lambda f + e)/r$, $f \sim$ Lognormal$(0,\Phi)$, $e \sim$ Lognormal$(0,\Psi)$, $r \sim \sqrt{\chi^2_5/3}$

[a] e, f and r are independent. T_{ML} is asymptotically valid only in condition I, all the other statistic are asymptotically valid for all the conditions.

Table 2
Marginal Skewness and Kurtosis

		Variables				
		(1 6 11)	(2 7 12)	(3 8 13)	(4 9 14)	(5 10 15)
II	β_1	0	0	0	0	0
	β_2	6	6	6	6	6
III	β_1	3.11	3.11	2.47	1.85	1.85
	β_2	90.22	90.22	67.98	47.97	47.97
IV	β_1	5.84	5.84	5.83	5.92	5.92
	β_2	159.98	159.98	159.91	166.97	166.97

to be identifiable, we restrict the last factor loading corresponding to each factor at its true value; this fixes the scale of the factors. All the other nonzero parameters are set as unknown free parameters. So $q = 33$ for this model, and $p^* - q = 87$.

The four distribution conditions as given in Table 1 were used in our study. Since $E(1/\chi^2_5) = 3$, the population covariances in conditions (I) to (IV) have the same covariance structure as given in Equation 7. So condition (I) generates data with a multivariate normal distribution, condition (II) generates data which is elliptically symmetric, condition (III) generates data which is asymmetrically distributed, but the common factor still follows a multivariate normal distribution before rescaling by $\sqrt{\chi^2_5/3}$; in condition (IV), both the common factors and unique factors are asymmetric in distribution. Table 2 gives the corresponding marginal skewnesses and kurtoses for the populations in conditions (II), (III) and (IV), based on

$$\beta_{1j} = \frac{E(x_j - \mu_j)^3}{[E(x_j - \mu_j)^2]^{3/2}} \quad \text{and} \quad \beta_{2j} = \frac{E(x_j - \mu_j)^4}{[E(x_j - \mu_j)^2]^2}.$$

Because of the symmetric nature of the design, the skewnesses and kurtoses of indicator variables for each factor are the same, and several variables have similar marginal parameters. For example, variables 1, 6, and 11 have the same skew and kurtosis, as do variables 2, 7, and 12. The distributional nature of the conditions is also reflected by these skewness and kurtoses, with condition II having no skew and moderate kurtosis, condition III having moderate skew and substantial kurtosis, and condition IV having more than moderate skew and extremely heavy kurtosis. Thus a wide range of skew and, especially, kurtosis is covered by the design. The empirical skew and kurtosis in the samples was consistent with the theoretical values of Table 2. The scaling factor $\sqrt{\chi^2_5/3}$ in conditions (II) to (IV) is to invalidate the asymptotic robustness condition of the normal theory method, so the statistic T_{ML} is only appropriate in condition (I). It is known that T_{SB} asymptotically follows a chi-square distribution with elliptical data in (II). Recent results by Yuan and Bentler (1997b) indicate that T_{SB} still asymptotically follows the nominal chi-square variate with nonnormal data as in (III) and (IV) even though these data are from distributions with heterogeneous marginal skewnesses and kurtoses. So statistics T_{SB}, T_B, T_{YB} and T_F are asymptotically valid for all the conditions. However, since our sample sizes are not large enough, these statistics can be expected to perform quite differently from each other. Hopefully, a single generally best-performing method will be found.

Based on degrees of freedom of p^* - q = 87 and the sample size requirement discussed earlier, we chose the following sample sizes for our study: N = 90, 100, 110, and 120 for T_B, T_{YB} and T_F. In order to see the performance of T_{SB} with rank $\left(\hat{U}S_Y\right)$ less than p^* - q, we also studied N = 60, 70, and 80. All of these sample sizes are below the minimum required for the ADF weight matrix S_Y to be positive definite, so we are investigating conditions when the classical ADF method cannot even be computed. Performance of these statistics for sample sizes above 150 can be found in Yuan and Bentler (1998).

For the performance of normal theory based inferences with normal and nonnormal data, T_{ML} was computed for all the sample sizes 60 to 120. With 500 replications, not all the 500 samples could reach a converged solution with the criterion $\left\|\theta^{(i+1)} - \theta^{(i)}\right\| \le 10^{-5}$ after 30 iterations. Our experience is that a converged solution would still not be reached even if we doubled the number of iterations. The results in Tables 3 to 6 (following pages) are based on the converged solutions, whose number N_c is given at the end of each table. We omit N_c if it equals 500. Specifically, we computed the rejection rate (R) based on 5% critical value from the corresponding nominal distribution of each statistic, and the sample means (M) and standard deviations (SD) of these statistics. It is known that for the nominal χ^2_{87}, $E(\chi^2_{87}) = 87$ and $\text{std}(\chi^2_{87}) = 13.19$. For the F-distribution, we also listed the population mean and standard deviation (in parenthesis) in Table 3 for easy reference. Note that for an F-distribution with 87 and 3 degrees of freedom, its population variance does not exist. So the sample standard deviation of the F-test corresponding to N = 90 may not be so informative.

Results

Results for the multivariate normal data are given in Table 3. Even though T_{ML} can still be obtained for all the sample sizes studied, it clearly does not behave like a nominal χ^2_{87} variate, especially for the smaller sample sizes. The statistic T_{SB} behaves even worse. For sample sizes between 60 and 80, we know that the matrix T_{SB} has rank less than p^* - q. As noted above, in such a case, the rescaling factor in T_{SB} is not strictly legitimate and we would not expect it to behave as a χ^2_{87} variate. However, it even behaves badly with sample sizes between 90 and 120 where in principle it could perform all right. It is already known that the statistic T_B behaves badly with small sample sizes, and these results verify this expectation. In fact, here performance is quite appalling at all the sample sizes. The statistic T_{YB} has a zero rejection rate for

Table 3
Performance of Five Statistics in Condition I Based on 500 Replications

		Sample Size						
		60	70	80	90	100	110	120
	R	114	101	79	72	71	67	59
T_{ML}	*M*	99.49	97.19	95.87	94.35	93.85	93.25	92.63
	SD	15.19	15.42	14.71	14.13	14.41	14.31	14.33
	R	145	134	100	87	80	80	63
T_{SB}	*M*	102.77	99.93	98.23	96.33	95.70	94.95	94.16
	SD	15.57	15.97	15.09	14.35	14.60	14.58	14.53
	R				500	500	500	500
T_B	*M*				10134.39	939.98	542.83	379.60
	SD				64313.50	454.89	200.59	119.89
	R				0	0	0	0
T_{YB}	*M*				87.42	88.90	89.90	89.66
	SD				1.95	4.04	5.54	6.63
	R				18	50	63	56
T_F	*M*				3.93 (3.00)	1.42 (1.18)	1.32 (1.10)	1.21 (1.06)
	SD				24.92 (*)	.687 (.591)	.487 (.400)	.382 (.326)

all the sample sizes studied here. This once again verifies that T_{YB} overcorrects the behavior of T_B for very small sample sizes. The statistic T_F performs best among all the statistics considered here, though it also has overrejections for sample sizes 100 to 120. Comparing the rejection rates with the sample mean and sample standard deviations of the corresponding statistics, we see that they are closely related. That is, a high rejection rate goes with either a large sample mean or a large sample standard deviation. For example, even though T_{YB} has a good approximation to the population mean, because its small sample standard deviation is way too low, its rejection rate also is too low.

Table 4 contains results for the elliptical data. The statistic T_{ML} rejects the true model almost all the time. It is not robust to elliptical data. This implies that invalidation of asymptotic robustness conditions also is relevant to small

Table 4
Performance of Five Statistics in Condition II Based on N_c Converged Replications

		Sample Size						
		60[a]	70[b]	80[c]	90[d]	100[e]	110[f]	120[g]
T_{ML}	*R*	483	486	478	487	487	485	486
	M	159.46	160.79	162.02	163.70	164.62	165.76	168.15
	SD	34.62	37.91	37.05	40.43	40.23	42.26	42.05
T_{SB}	*R*	148	118	85	66	60	58	46
	M	103.66	100.76	98.24	96.52	95.76	94.57	93.80
	SD	14.17	14.20	13.36	12.57	12.34	12.55	12.42
T_B	*R*				499	499	499	499
	M				7592.76	901.94	505.27	361.30
	SD				17230.92	421.10	170.42	102.29
T_{YB}	*R*				0	0	0	0
	M				87.50	88.55	88.97	88.83
	SD				1.91	4.07	5.28	6.14
T_F	*R*				27	43	42	35
	M				2.94	1.36	1.23	1.15
	SD				6.68	.636	.413	.326

[a]$N_c = 493$, [b]$N_c = 499$, [c]$N_c = 490$, [d]$N_c = 499$, [e]$N_c = 499$, [f]$N_c = 499$, [g]$N_c = 499$

samples. The statistic T_{SB} still overrejects samples, even though it works much better than T_{ML}. Statistics T_B and T_{YB} work similarly as they do with normal data. And finally, the statistic T_F still performs best among the alternative methods with elliptical data. In fact, it has a more accurate rejection rate here than it does in the case of multivariate normal data.

Tables 5 and 6 contain results for asymmetric data. Relatively speaking, these largely mirror results obtained under the elliptical condition. Of course there are specific differences. As compared with elliptical data, T_{ML} performs a little better here, but it still highly overrejects the true model, while T_{SB} performs somewhat worse in overrejections. The statistics T_B and T_{YB} perform similarly as compared to the elliptical data case. In contrast, the

Table 5
Performance of Five Statistics in Condition III Based on N_c Converged Replications

		Sample Size						
		60[a]	70[b]	80[c]	90[d]	100[e]	110[f]	120
T_{ML}	*R*	387	405	400	400	411	410	414
	M	142.53	144.06	141.54	146.08	145.88	145.35	147.30
	SD	38.89	41.26	40.51	45.55	42.13	42.83	43.64
T_{SB}	*R*	179	135	108	80	88	66	67
	M	105.63	102.54	99.64	98.80	98.30	96.66	95.75
	SD	13.62	13.37	12.92	12.77	12.39	12.25	12.56
T_B	*R*				496	499	499	500
	M				6624.79	848.31	491.65	361.45
	SD				10468.16	362.62	149.03	91.02
T_{YB}	*R*				0	0	0	0
	M				87.39	88.21	88.76	89.07
	SD				1.84	3.71	4.80	5.56
T_F	*R*				23	28	30	30
	M				2.57	1.28	1.19	1.15
	SD				4.06	.547	.361	.290

[a]$N_c = 490$, [b]$N_c = 496$, [c]$N_c = 491$, [d]$N_c = 496$, [e]$N_c = 499$, [f]$N_c = 499$

statistic T_F works best for asymmetric data. In these tables, its rejection rate can be seen to be very close to the target rate of 5%.

Conclusion and Recommendation

It is well known that T_{ML} is regularly used for model evaluation, especially when a data set is approximately normally distributed. This is an appropriate practice except under two conditions as studied here, namely, with nonnormal data and even with normal data when sample size is small. The statistic T_{SB}, which has been considered to work quite reliably under a wide

Table 6
Performance of Five Statistics in Condition IV Based on N_c Converged Replications

		Sample Size						
		60[a]	70[b]	80[c]	90[d]	100[e]	110[f]	120[g]
T_{ML}	*R*	377	360	376	398	379	384	396
	M	141.49	138.71	141.27	143.64	140.55	142.68	144.93
	SD	38.45	38.40	40.29	41.99	39.72	42.13	43.18
T_{SB}	*R*	179	123	97	103	78	69	62
	M	105.77	101.86	100.14	99.68	97.36	96.34	95.40
	SD	13.20	13.48	12.46	13.11	12.90	12.19	12.26
T_B	*R*				495	481	497	493
	M				7460.63	862.16	495.31	366.19
	SD				17466.47	403.28	155.82	89.95
T_{YB}	*R*				0	0	0	0
	M				87.45	88.44	88.85	89.43
	SD				1.81	3.47	4.82	5.32
T_F	*R*				24	21	32	29
	M				2.89	1.30	1.20	1.17
	SD				6.77	.609	.378	.287

[a]N_c = 482, [b]N_c = 476, [c]N_c = 483, [d]N_c = 495, [e]N_c = 481, [f]N_c = 497, [g]N_c = 493

variety of conditions (e.g., Hu et al., 1992; Curran et al., 1996), was found in this study to break down with smallest sample sizes under all conditions. At the larger of the small sample sizes, this statistic certainly did outperform T_{ML}, but it still rejected 2-4 times as many true models as it should have to yield nominal performance. For quite different reasons, statistics T_B and T_{YB} should not be used when sample size is smaller than the number of nonduplicated elements of the sample covariance. Browne's statistic essentially always rejects the true model. In contrast, Yuan and Bentler's statistic essentially always accepts the true model, when it should be at least occasionally rejecting this model by chance. In between these extremes, the new T_F

statistic performed remarkably well at all small sample sizes. Although it had some overrejection under conditions of normality, it still outperformed T_{ML} in this situation. Under conditions of nonnormality, it always yielded the best performance, especially with asymmetric data where it yielded remarkably close to nominal performance, even with the extremely nonnormal data of condition IV. Yuan and Bentler (in press) propose that the good performance of their *F*-test of model fit is most likely due to the general robustness properties of Hotelling's T^2 statistic, upon which it is based, to distributional violation.

In this article, we studied the various statistics with small sample sizes. Yuan and Bentler (1998) studied these statistics with medium to large sample sizes. Their key findings can be summarized as follows. They found that the statistic T_B gives reliable inference for the 15 variables factor model when sample size is above 5000, but it always overrejects the correct model for smaller sample sizes. On the other hand, T_{YB} overaccepts correct models for smaller sample sizes. For sample sizes of 200 and above, T_{YB} gives reliable inferences, and its performance is very stable under a variety of conditions. When sample sizes are greater than 200, the statistic T_{SB} gives very good performance if all the τ_j in Equation 4 are equal or nearly equal. When the τ_j are very different, the performance of T_{SB} tends to be worse as sample sizes get larger. Unfortunately, there is no practical procedure for checking the equality of the τ_j, though the coefficient of variation of the sample estimates $\hat{\tau}_j$ provides a plausible index. Finally, the statistic T_F behaves well for the 15 variables factor model, however, it tends to overreject correct models too

Table 7
Summary and Recommendation

Sample Size	Type of Data	Recommended Statistics
Small sample sizes		
$N \leq (p^* - q)$	Normal or nonnormal	Further study needed
$p^* - q \leq N \quad p^*$	Normal or nonnormal	T_F
Medium to large sample sizes $N > p^*$		
	Normal	T_{ML}
	Nonnormal	T_{YB} or T_F

often in a 15 variable intra-class model. Considering the results in Tables 3-6 and those of Yuan and Bentler (1998), our overall recommendations for practice are given in Table 7.

It must be recognized that our recommendations are based on limited experience with these various test statistics. The simulation work obviously covers only a specialized model under a very narrow range of conditions. Nonetheless, there is remarkable similarity in performance of the several test statistics across the various conditions. In fact, the results for normal data, shown in Table 3, already anticipate the main trends for various degrees of nonnormality up to the extreme kurtosis condition in Table 6. Whether or not the recommendations we give in Table 7 hold up under a greater variety of conditions such as variations in the number of variables, the type of model, and even violations of continuity, remains to be determined, but certainly no comfort should be taken in the poor performance of the oldest statistics T_{ML} and T_B. Although T_B has hardly ever been applied in practice, and while very little has been published about its performance, its almost universal rejection of the true model in this study is certainly cause for concern. This poor performance is even more extreme than the degraded performance of the standard ADF test with large models and intermediate sample sizes (e.g., Hu et al. 1992). Since the ADF test breaks down equally with continuous and categorical data (Muthén & Kaplan, 1992), we would expect T_B also to break down with typical Likert data. At the other end of the spectrum of use, it has been known for some time that the most widely used statistic T_{ML} also can perform very badly when data is not normally distributed (Hu et al., 1992; Curran et al., 1996), but this knowledge has not affected practice very much (Bentler & Dudgeon, 1996). The results in Table 3 indicate that this statistic misbehaves even with normal data when sample size is small, while Tables 4-6 show again that T_{ML} can be extremely misleading under nonnormality. Of course, we purposefully avoided asymptotic robustness conditions which conceivably might help T_{ML} perform better in the analysis of real data. But so far as we know, there is no evidence one way or the other on the relevance of asymptotic robustness theory to empirical multivariate data.

In this article, we studied the rejection rates of several test statistics. Another important aspect of performance is power, but the power of these statistics has not been addressed here and certainly requires study. Since theoretical results in this field have been asymptotic, the null and nonnull distributions of each statistic studied here can only be approximated in practice, and the accuracy of the approximation no doubt will depend on the type of data and sample size besides the correctness of the model. However, it is not difficult to predict that the statistic T_B will have a very high power to reject false models since its rejection rate is already 100% for correct models.

In contrast, the statistic T_{YB} most likely will have an attenuated power for smaller sample sizes. It is important to note that it is not difficult to invent a statistic that has a power of 1.0 by letting it reject any model (correct or wrong) or to invent a statistic that has zero type I error by letting it accept any model. A good statistic should possess the property of a controllable type I error while achieving a maximum power. Unfortunately, it is not as easy to control type I error and power in structural equation modeling as it is with simpler statistics such as the basic *z*-statistic.

Preliminary indications are that the two statistics recommended here for general purpose use, T_{YB} and T_F, as well as the more widely known T_{SB}, have good power for intermediate sample sizes. The power of the statistic T_{SB} was studied in Curran et al. (1996), with favorable though guarded results. The power of the other statistics based on an ADF estimator was studied in Yuan and Bentler (1997a, in press). If fitting the 3-factor model by a 2-factor model, the power of T_F and T_{YB} have average power of .847 and .559 at sample size 150, and.982 and.932 at sample size 200. So inference based on T_F and T_{YB} seems to be reliable with regard to power when sample sizes are above 200.

References

Amemiya, Y. & Anderson, T. W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *Annals of Statistics, 18*, 1453-1463.

Austin, J. T. & Calderón, R. F. (1996). Theoretical and technical contributions to structural equation modeling: An updated annotated bibliography. *Structural Equation Modeling, 3*, 105-175.

Bentler, P. M. & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology, 47*, 541-570.

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Browne, M. W. & Arminger G. (1995). Specification and estimation of mean and covariance models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185-249). New York: Plenum.

Browne, M. W. & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology, 41*, 193-208.

Curran, P. S., West, S. G. & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

Hu, L., Bentler, P. M. & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351-362.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183-202.

Magnus, J. R. & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. New York: Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Muthén, B. & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

Satorra, A. & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *American statistical association 1988 proceedings of business and economics sections* (pp. 308-313). Alexandria, VA: American Statistical Association.

Satorra, A. & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*, *10*, 235-249.

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Newbury Park, CA: Sage.

Yuan, K.-H. & Bentler, P. M. (1997a). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92*, 767-774.

Yuan, K.-H. & Bentler, P. M. (1997b). On normal theory and associated test statistics in covariance structure analysis under two classes of nonnormal distributions. Conditionally accepted by *Statistica Sinica*.

Yuan, K.-H. & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 51*, 289-309.

Yuan, K.-H. & Bentler, P. M. (in press). *F*-tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*.

Accepted June, 1998.